Representation, control, or reasoning? Distinct functions for theory of mind within the medial prefrontal cortex

Hartwright, Charlotte E; Apperly, Ian A; Hansen, Peter C

DOI:
10.1162/jocn_a_00520

Representation, Control, or Reasoning? Distinct Functions for Theory of Mind within the Medial Prefrontal Cortex

Charlotte E. Hartwright, Ian A. Apperly, and Peter C. Hansen

Abstract

■ The medial pFC (mPFC) is frequently reported to play a central role in Theory of Mind (ToM). However, the contribution of this large cortical region in ToM is not well understood. Combining a novel behavioral task with fMRI, we sought to demonstrate functional divisions between dorsal and rostral mPFC. All conditions of the task required the representation of mental states (beliefs and desires). The level of demands on cognitive control (high vs. low) and the nature of the demands on reasoning (deductive vs. abductive) were varied orthogonally between conditions. Activation in dorsal mPFC was modulated by the need for control, whereas rostral mPFC was modulated by reasoning demands. These findings fit with previously suggested domain-general functions for different parts of mPFC and suggest that these functions are recruited selectively in the service of ToM. ■

INTRODUCTION

Theory of Mind (ToM) is a term used to describe the ability to attribute mental states such as beliefs, desires, and intentions to other individuals. By applying a ToM, social agents are better able to predict the behavior of those around them and may additionally direct our own behavior in terms of whether we choose to deceive, cooperate, or empathize with others (Gallagher & Frith, 2003). This ability to "mentalize" has received much attention from the neuroimaging community over the last decade and has identified a set of brain regions that are consistently responsive when thinking about the contents of other people's minds: the left and right TPJ, medial parietal cortices including the precuneus and posterior cingulate, the temporal poles, and the medial pFC (mPFC; for reviews, see Mar, 2011; Carrington & Bailey, 2009; Van Overwalle, 2009; Lieberman, 2007). The most prominent debate within the literature, however, surrounds how medial prefrontal and temporoparietal regions support a functioning ToM.

One challenge for social neuroscientists is to localize ToM processes more precisely by identifying functional subdivisions within the anatomical regions associated with ToM. mPFC, in particular, comprises a large area of the cortex and is involved in many aspects of social cognition (Amodio & Frith, 2006), together with an array of executive processes such as reallocation of attention, action monitoring and control (Lieberman, 2007; Rushworth, Buckley, Behrens, Walton, & Bannerman, 2007; Ramnani & Owen, 2004), relational integration and multitasking (Gilbert et al., 2006; Ramnani & Owen, 2004), outcome monitoring (Gilbert et al., 2006), working memory and episodic memory (Spreng, Mar, & Kim, 2009; Lieberman, 2007; Gilbert et al., 2006; Ramnani & Owen, 2004) and default mode or spontaneous "at rest" cognition (Spreng et al., 2009; Amodio & Frith, 2006; Ramnani & Owen, 2004). Because ToM is a social process but undoubtedly also entails executive processing, attention, and reasoning (Apperly, 2010), it is perhaps unsurprising that the role of mPFC in ToM remains unclear (Rothmayr et al., 2011).

It is likely that activation of mPFC, in some ToM tasks, reflects executive processes that are an incidental feature of the task used to present the ToM problem. Thus these do not constitute core processes that underlie ToM. Nonetheless, there are also good reasons for believing that specific subregions of mPFC are more centrally involved in ToM. A task analysis of ToM suggests three processes that may explain how specific regions of mPFC are involved in mentalizing. First, a common theme across all forms of ToM reasoning is the requirement for representation of a mental state. Thus, regardless of whether an individual is asked to reason about an agent's belief, desire, intention, or the like, it is necessary to represent a mental state of some kind. The frequency with which more rostral areas of mPFC are recruited for ToM and other social cognitive functions has led researchers to tentatively suggest that mPFC might subserve such a process (Amodio & Frith, 2006; Frith & Frith, 2003, 2006). Other data suggest that TPJ may be even more selectively responsive than mPFC to representation of mental states (Aichhorn et al., 2009; Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, &

University of Birmingham, UK

Journal of Cognitive Neuroscience 26:4, pp. 683–698
doi:10.1162/jocn_a_00520

Saxe, 2009; Saxe & Wexler, 2005; Saxe & Kanwisher, 2003). Resolution of this debate is unnecessary for our current purposes. What matters for now is that there are grounds to suppose that mPFC may be involved in representation of mental states and that it is possible to distinguish this representational requirement, which attends all ToM tasks, from other important requirements for cognitive control and reasoning, which vary across ToM tasks or experimental conditions.

Second, a large body of behavioral and neural evidence indicates that ToM is associated with processes for cognitive control (e.g., Apperly, 2010; Lieberman, 2007). Control processes for inhibition, conflict monitoring, and working memory are not only necessary for meeting the demands of the relatively complex stories or cartoons frequently used to study ToM but also seem to be essential for ToM per se. For example, in the classic false belief paradigm (see Wimmer & Perner, 1983), an agent holds an outdated or "false" belief about reality. Predicting the agent's action requires participants first to infer that the agent's belief is different from their own, second to hold this false belief in mind and not confuse it with their own knowledge, and third to predict the agent's action selectively on the basis of the agent's belief, rather than according to the participant's own knowledge of the right course of action. Behavioral data from both children and adults suggest that the effort required for ToM reasoning (as indexed by response times and error rates) depends on whether an agent's belief is true or false and whether their desire is to approach or avoid a target object (Hartwright, Apperly, & Hansen, 2012; Apperly, Warren, Andrews, Grant, & Todd, 2011; German & Hehman, 2006). Attempts to understand the neural basis of such effort consistently identify more dorsal regions of mPFC (dmPFC) approximating BA 8, BA 9, and BA 32. For example, dmPFC is modulated by contrasting ToM concepts where maximal conflict exists, as is the case in false belief reasoning versus reasoning about an agent whose belief is a "true" representation of reality (Döhnel et al., 2012; Hartwright et al., 2012; Sommer et al., 2007).

Importantly, recent evidence suggests that the contribution of frontal regions does not just vary according to overall task difficulty, but according to the source of that difficulty in the ToM task. In a recent study on which the present paradigm is based, participants predicted the action of an agent whose belief was either true or false and whose desire was either to approach or avoid an object (Hartwright et al., 2012). Both factors have the potential to vary cognitive conflict, because both false belief and avoidance desire lead the agent into counterintuitive actions away from a salient target object. However, only the belief factor (true vs. false) leads to systematic variation in perspective between the character and the participant. In this study, dmPFC was modulated equally by the belief and desire factors, suggesting that it was performing a general role in resolving cognitive conflict. This contrasted with more lateral prefrontal regions, such as bilateral inferior frontal gyrus, which responded differentially to true versus false belief, but not to approach versus avoidance desire. These findings suggest that dmPFC underlies frontline control processes, which monitor conflict during ToM reasoning, whereas frontolateral regions, such as inferior frontal gyrus, are recruited for more specific processes such as inhibition of self-perspective. This theory about the contribution of dmPFC converges with neuroimaging research outside the social domain, which identifies mPFC, particularly more dorsal regions including the dorsal ACC, in conflict monitoring and error detection. It has been shown that activation in dmPFC is modulated by task difficulty, where those tasks that attract increased error rates and response latencies make the most demands on this region (Botvinick, Cohen, & Carter, 2004; Bush, Luu, & Posner, 2000; Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999).

The final process we propose within our task analysis of ToM is the focus of the current study and refers to the different roles of reasoning. Philosophers, logicians, and computational scientists have long debated the formulation of reasoning. These debates are beyond the scope of the present article; however, we borrow two theoretical concepts to illustrate how different approaches to ToM can activate alternative modes of inference and their neural correlates. In the belief desire task used by Hartwright et al. (2012), participants were told three facts for each trial: the agent's belief about the location of an object, the agent's desire to seek out or avoid the object, and the true location of the object. Given this information, participants had to identify which location the agent would choose on the basis of his belief and desire state. Thus, participants had to reason "deductively," so no reasoning beyond the facts explicitly presented was required (Pagnucco, 1996; Morris, 1992). Unlike the vast majority of neuroimaging studies of ToM, activation within rostral mPFC (rmPFC), approximating BA10, was noticeably absent from this deductive ToM paradigm.

However, many ToM studies—and certainly a good deal of ToM outside the laboratory—do not provide explicit access to all of the facts necessary to solve the task (Jenkins & Mitchell, 2009). Consequently, the individual is required to engage in open-ended "abductive" reasoning about an agent's behavior to use their ToM effectively. Consider a typical ToM vignette taken from Saxe and Andrews-Hanna (n.d.),

> The morning of the high school disco Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed.
>
> -
>
> Sarah gets ready assuming her shoes are under her dress.
>
> TRUE/FALSE

Unlike the deductive approach, reasoning here is used to explain an observation on the basis of a hypothesis, which may or may not turn out to be correct. Here,

participants are required to reason abductively—that is, to infer the most likely cause (Sarah's belief that her shoes are under her dress) on the basis of the given effect (that Sarah gets ready unaware that her shoes might not be where she expects to find them) and a ToM principle (that Sarah will look for the shoes on the basis of her belief state). Here, then, reasoning is an inference to the most appropriate explanation (Menzies, 1996). Reasoning deductively, where one uses a set of rules and preconditions to generate a conclusion (Menzies, 1996), is likely to involve cognitive process that differ from an abductive approach involving reasoning to explain an observation (Morris, 1992). While the neural basis of deductive reasoning has been studied extensively (see Prado, Chadha, & Booth, 2011, for a recent review), little work has been done for abductive inference. Nonetheless, when considered in terms of the underlying process of thinking beyond the given information, studies indicate that rmPFC is recruited when participants are required to reason beyond the constraints of the information immediately available to them (Hartwright et al., 2012; Jenkins & Mitchell, 2009; Gilbert et al., 2007), whether the context is social or nonsocial. This leads to the hypothesis that, rather than being involved in representing mental states, rmPFC is recruited whenever ToM tasks require abductive reasoning. This would account for the frequent observation of rmPFC activation because abductive reasoning is very common in ToM tasks.

However, there is an alternative explanation for the lack of variable mPFC activation in Hartwright et al. (2012) that remains consistent with the hypothesis that rmPFC supports the representational demands of ToM, as touched upon earlier in our task analysis. The need to represent mental states was present across all conditions in Hartwright et al.'s deductive task; consequently, rmPFC might not be identified by orthogonal comparisons across conditions if this region generally services the process of representation. Therefore, this study manipulates the need for abductive reasoning within task to disambiguate these two possibilities.

In summary, there are multiple theoretical reasons for thinking that mPFC might be involved in ToM and several competing hypotheses designed to account for this. Furthermore, there are grounds for thinking that there might actually be functional differentiation within mPFC, which a number of researchers have suggested would be best identified using a single, within-experiment, within-subject design (Abu-Akel & Shamay-Tsoory, 2011; Carrington & Bailey, 2009). The task analysis presented here proposes three separate processes for ToM: representation, control, and reasoning. Representation, we argue, is a ubiquitous feature of mentalizing. Control and reasoning processes, conversely, vary across different ToM tasks. The latter two ToM processes lend themselves well to manipulation within a single, repeated-measures paradigm. Consequently, this study served two purposes. First, to replicate Hartwright et al.'s earlier finding that dmPFC is modulated as a function of control. Second, by making a minimal change to our previous paradigm, we aimed to demonstrate that we could recruit the previously absent rmPFC by including a condition that required abductive reasoning. To achieve this, we present a 2×3 repeated-measures orthogonal design. The valence of an agent's belief was either true or false; the valence of desire was either approach or avoidance (as in Hartwright et al., 2012) or it was unspecified. The novel, unspecified, condition required participants to reason about whether they thought the agent would have an approach or an avoidance desire, on the basis of what sort of person they thought the agent was. We expected those mental states where conflict is inherent but presented unambiguously in our paradigm (i.e., false belief, avoidance desire), to preferentially recruit dmPFC. Conversely, a mental state that required abductive reasoning (i.e., desire unspecified) was expected to preferentially activate rmPFC.

METHODS

Participants

Twenty right-handed adults participated in the fMRI experiments (12 women; mean age = 21 years). All were native English speakers and were given a small honorarium for their participation. The study had research ethics approval from the University of Birmingham. All participants gave written consent to participate in the study.

Prescreen

A prescreen to determine suitability to participate was conducted several days before collecting any neuroimaging data. This consisted of a handedness measure, using a modified form of the Annett Handedness Questionnaire (1970), and a reading scale—the Wide Range Achievement Test Third Edition—to ensure reading proficiency commensurate with the experimental tasks.

Participants were informed that the social judgments task required them to make predictions about how real individuals played a game in a previous experiment. They then completed a computer-based interactive training session and two test blocks of the task. Those who performed above chance on the test blocks were invited to participate in the fMRI experiment.

Social Judgments Experiment

The social judgments task was based on a paradigm devised by Hartwright et al. (2012) and Apperly et al. (2011). The experiment comprised an orthogonal design where a protagonist's belief state (true [B+] or false [B−]) and desire state (approach [D+], avoid [D−], or unspecified [D±]) was systematically manipulated, resulting in six equally occurring conditions: B+D+, B+D−,

B+D±, B−D+, B−D−, B−D±. Immediately before collecting any neuroimaging data, participants were again informed that the task was based on real game playing data from real individuals and that the participant's job was to predict how these individuals played the game. All participants then revisited the interactive training program used in the prescreen and completed a further practice block outside the MRI scanner. Note that none of the practice trials were used in the fMRI experiment.

The fMRI experiment required the participants to watch and predict which one of two different colored boxes a character, referred to as "the contestant," would open in a virtual game show (Figure 1). A single round (i.e., trial) of the game show consisted of the contestant being told what prize was on offer, followed by them guessing which one of the two boxes contained the prize, ending in them opening one box. The contestant would win whatever was in the box they opened; however, one box was always empty and the other always contained the prize. If the box contained a prize, they would win it. If it was empty, they would win nothing and play a new round.

Each contestant played multiple rounds. The prizes ranged in desirability, and as the contestant could only win a finite number of prizes, it was not always in their interest to play to win every prize. If the contestant liked the prize on offer, they would open the box where they guessed the prize was hidden, in the hope of winning that prize. If they did not like the prize, they would open the opposite box to where they guessed the prize was hidden (i.e., the empty box), in the hope of having another chance to win something more to their liking. Note, however, that the game show was designed such that the contestant would only take home a prize in half of the trials. Furthermore, in half of the "winning" trials, the contestant would win a prize that they did not actually want to win.

While the fMRI data were collected, participants watched a computer-based mock-up of the contestants playing the afore-described game show. The participants' job was to predict which box the contestant opened on the basis of the contestant's belief and desire state, in terms of which of the two boxes the contestant believed contained the prize, and the contestant's desire to win or gamble and play on for a better prize. Participants were always told the contestant's belief about the location of the prize and the true location of the prize, but had to infer the contestant's desire to win the prize based on a color photograph that depicted the contestant smiling (D+), frowning (D−), or with a neutral (D±) expression. The training sessions conducted before collecting any fMRI data taught the participants to treat a smiling face as signaling the contestant's pleasure and, therefore, their desire to open the box that they thought contained the prize (approach desire) and a frowning face as signaling the contestant's displeasure and, therefore, their desire to avoid opening the box that they thought contained the prize (avoidance desire). Where the contestant was shown with a neutral expression, participants were asked to consider what sort of person they thought the contestant was, in terms of what their likes and dislikes might be, and to select which box they thought the contestant would open (unspecified desire). Just as with approach/avoid (D+/D−) trials, in these unspecified desire (D±) trials,

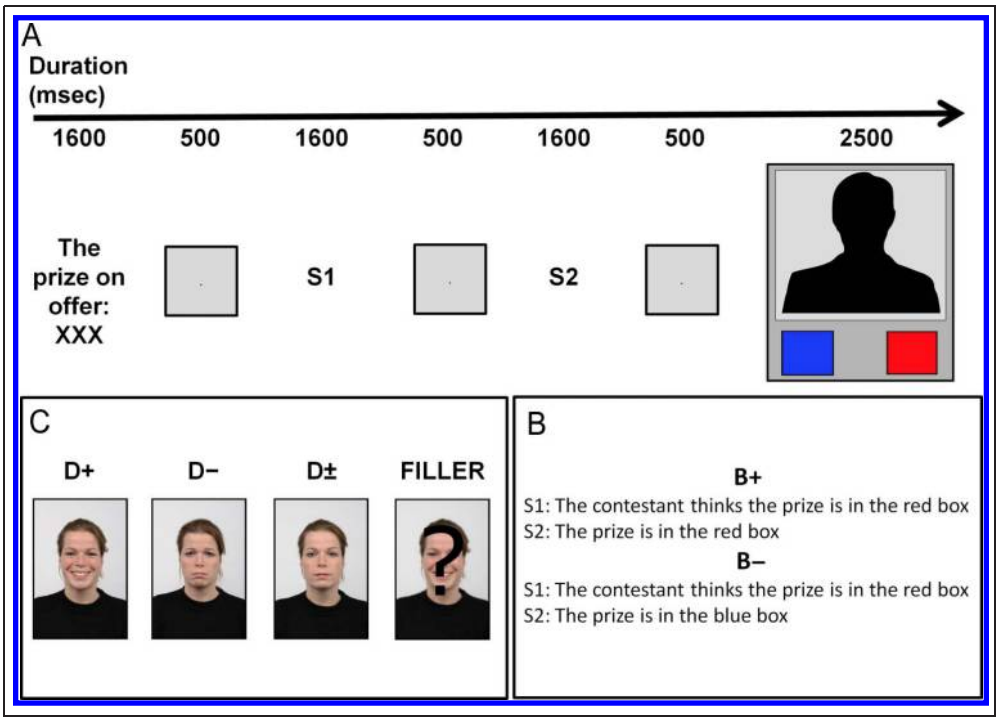

Figure 1. (A) Schematic example of a single trial. The left/right presentation of the red/blue box was randomized. Where XXX is written for the prize on offer; this would name a unique item for each trial, e.g., The prize on offer: hot tub. (B) Example statements for true (B+) and false (B−) belief scenarios. The temporal order of these statements was randomized. (C) From left to right, example response probe for approach (D+), avoidance (D−), unspecified desire (D±), and filler trials.

participants were told to select the box that the contestant believed contained the prize if they thought the contestant would have played to win the prize on offer, or to select the opposite box (i.e., what the contestant believed to be the empty box) if they thought the contestant would have wanted to avoid winning the prize. Note that in all cases, participants were told to make their responses on the basis of the contestant's belief state, which could be either true (B+) or false (B−), therefore requiring them to ignore their own knowledge of the true location of the prize.

Each block of trials opened with an instruction screen followed by an initial ISI of 11,600 msec. A single trial comprised three center-justified statements shown singularly for 1600 msec and separated by a 500-msec fixation period, followed by a picture response probe shown for 2500 msec, then a rest period. A variable ISI was used for rest (range = 9000–14,000 msec, $\bar{X}$ = 11,500 msec) during which a small fixation dot was displayed. Each trial lasted 8800 msec, excluding fixation. The experiment comprised six separate blocks, each of which contained 28 trials and took 9 min 36 sec to complete.

Each trial opened with a prize statement (e.g., The prize on offer: designer shoes), followed by either a belief statement (e.g., The contestant thinks the prize is in the red box) or a reality statement (e.g., The prize is in the blue box), then the remaining belief or reality statement. The temporal order of belief and reality statements was randomized but contained an equal number of each ordering overall. The final statement was followed by a response probe then rest. Participants were able to respond from the onset of the response probe, using a two button box placed in their right hand. Participants responded by pressing the left button to indicate the left prize box and the right button to indicate the right prize box.

Two formats of response probe were used. The format indicated to the participant what type of response to give. If a full color photograph of the contestant was shown, the participant had been trained to indicate which box they thought the contestant opened, based on the contestant's belief desire state. These were the trials of interest and made up 75% of the total number of trials. To ensure that the participants must attend to the contestant's belief state, regardless of whether it was true or false, antistrategy trials, termed herein "fillers," formed 25% of the presented trials (see Hartwright et al., 2012, for further discussion). Here, the response probe consisted of a full color photograph of the contestant, which had been blurred using a Gaussian smoothing kernel of 10 pixels FWHM. A black question mark obscured part of the contestant's face. Participants had been trained to indicate the true location of the prize when this format of response probe was shown. These fillers did not form any part of the analyses presented here.

Images of the contestants were taken from the Radboud Faces Database (Langner et al., 2010). Twenty-eight contestant's featured in the experiment (all white; 14 men), where each face was shown on six occasions throughout the experiment, once per block. Each facial expression—happy, sad, neutral—was shown twice for each face. Each image consisted of a head and shoulders shot on a plain gray background. All contestants were wearing a plain black t-shirt. Each participant viewed a total of 168 rounds of the game show, made up of 126 trials of interest and 42 antistrategy fillers, each with a unique prize, presented over the six blocks.

DATA ACQUISITION

Data were acquired in a single session using a 3T Philips Achieva scanner, with an eight-channel head coil. Whole-brain coverage was achieved with the following parameters: repetition time = 2.5 sec, echo time = 35 msec, acquisition matrix = 96 × 96, flip angle = 83°, SENSE factor = 2. 232 T2*-weighted EPI volumes were obtained per block of the experiment, each of which consisting of 42 axial slices obtained consecutively in a bottom–up sequence, reconstructed voxel size = 3 × 3 × 3 mm^3. Four dummy volumes were acquired at scan time; these were removed before image reconstruction. Following acquisition of the functional data, a T1-weighted anatomical image was acquired (3-D TFE, sagittal orientation, repetition time = 8.4 msec, echo time = 3.8, matrix size 288 × 288, 175 slices, reconstructed voxel size = 1 × 1 × 1 mm^3). During the acquisition of functional data, Presentation software (v. 14.1; Neurobehavioral Systems, Albany, CA) was used to display the stimuli and record the behavioral response data simultaneously.

WHOLE-BRAIN ANALYSIS

The FMRIB software library (FSL version v.5.98; FMRIB, Oxford, www.fmrib.ox.ac.uk/fsl) was used to perform all preprocessing and statistical analyses. Preprocessing of the functional data consisted of slice timing (regular up) and motion correction (MCFLIRT). High-pass filtering was conducted on the BOLD signals using a Gaussian weighted filter of 30 sec. Spatial smoothing was then applied using a 5-mm FWHM kernel. The functional data were registered to their respective structural images and transformed to the Montreal Neurological Institute (MNI) reference brain using a 7-DoF linear transformation (FLIRT).

The modeling approach replicates the procedure outlined in Hartwright et al. (2012), which allows direct comparison following the minimal change to our previous paradigm. Six explanatory variables (EVs) of interest—B+D+, B+D−, B+D±, B−D+, B−D−, B−D±—were modeled to reflect the six experimental conditions. The onset of each EV was time-locked to the button response and reflected an arbitrary duration of 100 msec. Because of anticipated differences in RTs as a function of experimental condition, this approach ensured that activation

reflected the decision-making phase within the experimental sequence. This approach mirrors Hartwright et al. (2012), which was adopted following careful inspection of time series data. Each EV was convolved with a gamma-derived hemodynamic response function within a general linear model framework. The time series before the onset of the response probe was modeled as a regressor of no interest and orthogonalized with respect to the main EVs. Motion parameters and filler trials were also modeled as regressors of no interest. Higher- level modeling was used to aggregate the data across participants within a mixed effects model using cluster-based thresholding at voxel $Z > 2.5$, cluster $p_{corr} < .001$. Note that this particular threshold was applied for ease of comparison with the earlier published version of this paradigm. This final whole-brain result reflected a 2 × 3 repeated-measures ANOVA with Belief (B+/B−) and Desire (D+/D−/D±) as within-subject factors, plus eight contrasts for directional tests comparing the levels of each main factor (B+ > B−, B− > B+, etc.).

CONTRAST MASKING ANALYSIS

To demonstrate voxels that were preferentially active for each of the three levels within the factor of Desire, using FSL's command line tools (fslmaths), the corrected, thresholded data from the directional whole-brain analysis were used as inputs to generate three masks, $D+_{pref}$, $D-_{pref}$, and $D\pm_{pref}$. This was done by computing a logical AND, which collapsed across the pairs of directional contrasts for each level (i.e., $D+_{pref}$ = D+ > D− AND D+ > D±; $D-_{pref}$ = D− > D+ AND D− > D±; $D\pm_{pref}$ = D± > D+ AND D± > D−). Note that this analysis was not required for the factor of Belief, as the directional contrasts serve this purpose ($B+_{pref}$ = B+ > B−; $B-_{pref}$ = B− > B+). The mean effect across all conditions was also computed for significantly active voxels within a bisected ROI (slices X = −10 through to +10; MNI coordinates). This enabled identification of voxels that were preferentially active for unspecified desire (D±) versus all other belief (B+/B−) and desire (D+/D−) conditions within mPFC.

RESULTS

Behavioral Data

All RTs were recorded from the onset of the response probe. Any incorrect responses were removed for RT analysis. Note that correct responses are only applicable in D+/D− trials as D± requires a subjective judgment. A 2 × 3 repeated-measures ANOVA was conducted on the remaining RT data, with Belief (B+/B−) and Desire (D+/D−/D±) as within-subject factors. This revealed significant main effects of Belief, where B− > B+, $F(1, 19) = 166.65$, $p < .001$, $\eta^2 = 0.90$, and Desire, where D± > D− > D+, $F(1, 19) = 99.40$, $p < .001$, $\eta^2 = 0.84$, and a significant interaction, $F(2, 38) = 4.86$, $p < .05$, $\eta^2 = 0.20$. Simple effects analyses revealed significant effects of Belief at each level of the factor of desire, and significant effects of Desire at the two levels of the Belief factor (all ps < .01), however, with the interaction being accounted for by the effect of Belief being largest when desires were negative. A further two-way ANOVA was computed on the error data, with Belief (B+/B−) and Desire (D+/D−) as repeated measures. This identified a main effect of Belief, where errors B− > B+, $F(1, 19) = 6.68$, $p < .05$, $\eta^2 = 0.26$, and Desire, where errors D− > D+, $F(1, 19) = 8.35$, $p < .01$, $\eta^2 = 0.31$. No interaction was identified, $F(1, 19) = 0.20$, $p = .66$, $\eta^2 = 0.01$. Figure 2 summarizes the mean RT (A) and accuracy data (B).

fMRI Data

Whole-brain Analysis

A 2 × 3 repeated-measures ANOVA identified main effects of Belief (B+/B−) and Desire (D+/D−/D±) but no interaction between the two factors. Manipulation of an agent's belief state replicated our previously published findings (Hartwright et al., 2012), yielding regions regularly implicated in ToM such as bilateral TPJ and precuneus.

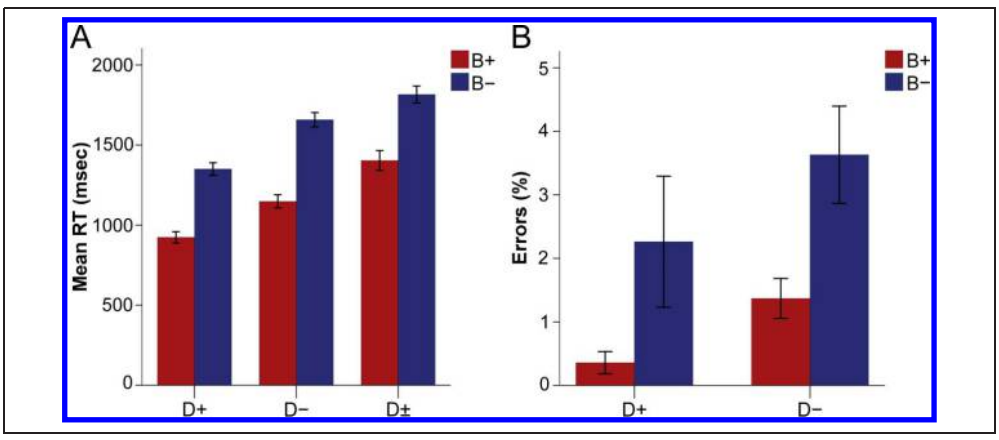

Figure 2. Error bars reflect ±1 *SEM*. (A) Group mean RT per condition for correct responses (msec): B+D+ = 923.57; B+D− = 1147.77; B+D± = 1403.10; B−D+ = 1350.18; B−D− = 1657.81, B−D± = 1815.55. Error bars reflect ±1 *SEM*. (B) Group mean percentage of errors across the four conditions where error data could be obtained: B+D+ = 0.36%; B+D− = 1.37%; B−D+ = 2.26%; B−D− = 3.63%.

Table 1. Factorial Analysis of Belief and Desire

Region	*Hemi*	*Brodmann's area*	*MNI Coordinates*			*Z Value*
			x	*y*	*z*	
Main Effect of Belief						
Temporoparietal junction	R	22	54	−56	26	6.47
Precuneus cortex	R	7	2	−66	48	5.76
Orbital frontal cortex	L	47	−32	26	−2	5.42
Temporoparietal junction	L	40	−52	−52	32	5.41
Insular cortex	R	47	46	16	−6	5.22
Middle frontal gyrus	R	44	50	20	38	5.18
Paracingulate gyrus	L	8	−4	20	48	5.09
Frontal pole	R	46	38	52	18	5.03
Inferior frontal gyrus, pars opercularis	L	45	−48	16	0	4.97
Precuneus cortex	L	7	−6	−66	54	4.94
Paracingulate gyrus	R	32	2	42	28	4.86
Middle frontal gyrus	L	44	−46	14	36	4.80
Insular cortex	L	47	−40	16	−6	4.40
Superior frontal gyrus	R	9	2	40	42	4.38
Inferior frontal gyrus, pars opercularis	R	48	52	18	4	4.34
Supramarginal gyrus, anterior division	L	40	−54	−40	38	4.21
Lateral occipital cortex, superior division	R	39	44	−60	52	4.18
Supramarginal gyrus, posterior division	R	40	48	−46	42	4.12
Postcentral gyrus	L/R	5	0	−54	72	3.52
Superior frontal gyrus	L	6	−2	14	68	3.29
Cingulate gyrus, anterior division	L	32	−8	40	16	3.16
Main Effect of Desire						
Paracingulate gyrus	R	8	2	24	48	8.34
Paracingulate gyrus	L/R	8	0	28	42	8.15
Occipital pole	R	18	22	−98	−2	7.42
Occipital pole	L	18	−24	−94	−8	7.16
Superior frontal gyrus	L	8	−8	30	58	6.34
Orbital frontal cortex	R	47	36	24	−6	5.78
Occipital fusiform gyrus	R	18	18	−84	−8	5.66
Intracalcarine cortex	R	17	14	−84	2	5.63
Superior frontal gyrus	R	8	4	54	40	5.61
Occipital fusiform gyrus	L	18	−14	−84	−12	5.59
Middle frontal gyrus	R	45	52	28	24	5.57
Lateral occipital cortex, superior division	L	19	−22	−86	20	5.56
Inferior frontal gyrus, pars opercularis	R	48	54	20	6	5.50
Orbital frontal cortex	L	47	−38	22	−8	5.29

Table 1. (*continued*)

Region	Hemi	Brodmann's area	MNI Coordinates			Z Value
			x	*y*	*z*	
Frontal pole	R	46	24	56	22	5.07
Cingulate gyrus, anterior division	L/R	32	0	44	14	5.01
Temporoparietal junction	L	39	−46	−58	44	4.70
Inferior frontal gyrus, pars triangularis	L	45	−50	20	0	4.43
Temporal pole	L	38	−48	16	−10	4.28
Inferior frontal gyrus, pars triangularis	R	45	54	34	12	4.19
Supramarginal gyrus, posterior division	R	40	52	−44	48	4.04
Frontal pole	L	47	−50	34	−20	4.03
Temporoparietal junction	R	22	60	−58	26	3.96
Middle frontal gyrus	L	44	−48	16	36	3.92
Supramarginal gyrus, anterior division	L	40	−42	−38	38	3.88
Supramarginal gyrus, posterior division	L	40	−42	−44	38	3.88
Inferior frontal gyrus, pars opercularis	L	45	−52	22	22	3.69
Temporal pole	L	38	−40	28	−24	3.62
Lateral occipital cortex, superior division	R	39	54	−62	34	3.54
Supramarginal gyrus, anterior division	R	2	54	−32	48	3.43
Postcentral gyrus	L	40	−32	−36	42	2.59

Regions identified using *F*-contrasts in a two-way repeated-measures ANOVA with factors of Belief (B+/B−) and Desire (D+/D−/D±). Table lists local maxima for cortical regions, which are modulated by varying belief status (true/false) and desire status (approach/avoid/unspecified) $Z > 2.5$, $p_{corr} < .001$. All anatomically unique local maxima (with minimum peak separation of 5 mm) are listed. Brodmann's areas are approximate.

Variation of an agent's belief state modulated considerable portions of the frontal cortex, including bilateral dorsolateral and ventrolateral prefrontal cortices spanning middle frontal and inferior frontal gyri, extending to orbital frontal cortex. This factor also recruited bilateral dorsal medial frontal regions comprising superior frontal, dorsal anterior cingulate, and dorsal paracingulate gyri (Table 1; red shading in Figure 3). Similar to belief reasoning, manipulation of an agent's desire state also recruited bilateral TPJ. However, lateral and medial prefrontal regions were recruited more extensively; thus, encompassed bilateral frontal poles on the lateral and medial surface, as well as rostral medial frontal regions including more ventral sections of the anterior cingulate and paracingulate gyri. Unlike the belief condition, variation of an agent's desire state also saw extensive recruitment of occipital regions spanning bilateral occipital poles to anterior occipital regions such as the calcarine cortex (Table 1; green shading in Figure 3).

Directional and Contrast Masking Analysis

A series of directional contrasts (Table 2) demonstrated that bilateral TPJ, superior parietal and occipital cortices, plus lateral and dorsal medial frontal regions, were typically more responsive when applying false over true belief reasoning to an agent (Figure 4A, $B-_{pref}$). The only regions that were preferentially active for true over false belief reasoning were the left occipital pole and occipital cortex (Figure 4A, $B+_{pref}$). For desire-based reasoning, contrast mask analyses indicated that bilateral occipital cortices and bilateral pre- and post-central gyri were preferentially responsive when applying an approach versus avoidance or unspecified desire (Figure 4B, $D+_{pref}$). When the agent expressed an avoidance versus an approach or unspecified desire, right precuneus was the only region to be preferentially recruited (Figure 4B, $D-_{pref}$). A large area covering medial and lateral pFC was highlighted to be most responsive when the agent's desire was unspecified versus to approach or avoid. Medial frontal activation spanned anterior cingulate, dorsal and rostral medial prefrontal cortices, extending laterally to bilateral frontal poles (Figure 4B, $D\pm_{pref}$). As shown in Figure 4C, rostral mPFC was preferentially active for unspecified desire over and above all of the other belief and desire states.

DISCUSSION

The diverse array of social and nonsocial tasks that activate mPFC has meant that the precise role of this region in ToM

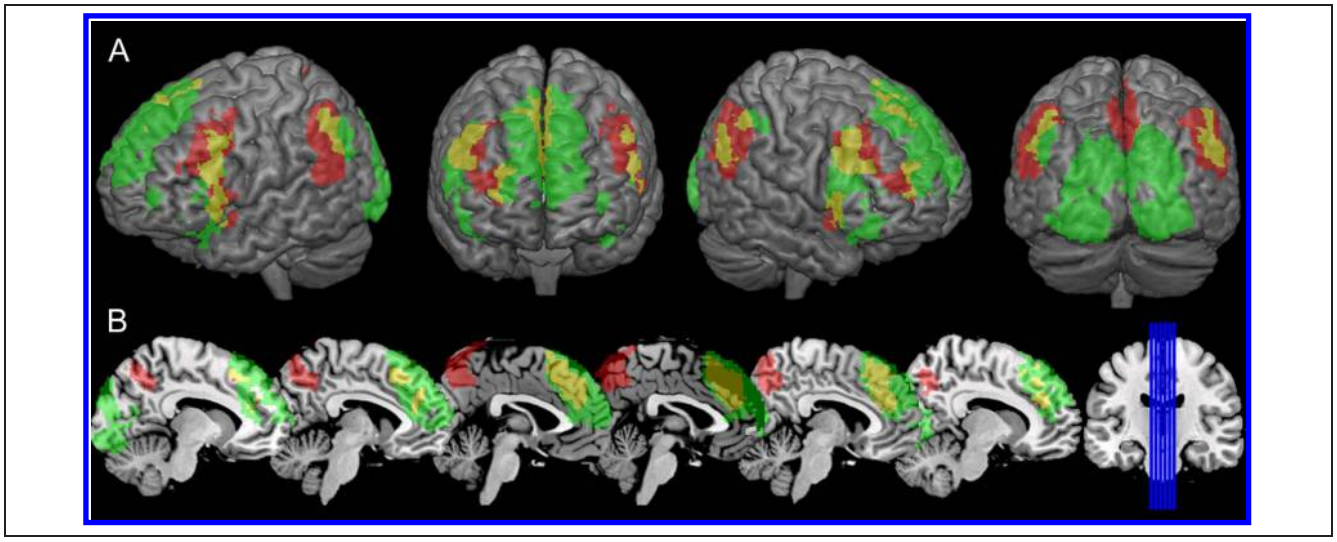

Figure 3. Result from 2 × 3 repeated-measures ANOVA whole-brain analysis with Belief (B+/B−; red) and Desire (D+/D±/D−; green) as within-subject factors. Yellow areas indicate regions recruited by both factors (B/D). The group data are overlaid on the MNI brain template, showing significantly activated voxels where $Z > 2.5$, $p_{corr} < .001$. Maps reflect *Z*-corrected *F*-statistical images and are displayed in neurological convention, where left is represented on the left side of the image. (A) Activation maps highlighting modulation on the lateral surface. Images from left to right show left, anterior, right, and posterior views of the cortex respectively. (B) Selected slices highlight modulation in medial frontal regions. Slices from left to right, $x = -10, -6, -2, 2, 6, 10$.

has remained vague (Rothmayr et al., 2011). We employed an analysis of common features of ToM tasks to distinguish roles that mPFC might serve for representation, control, and reasoning. The need to represent mental states was present in all task conditions, whereas the task made it possible for the first time to manipulate demands on control and reasoning within a single study. Our results suggest that dorsal and rostral regions of mPFC play distinctive roles in ToM control and ToM reasoning, respectively, and that these patterns are consistent with the proposed functions of these regions in nonsocial tasks.

Conflict Monitoring, Control, and the dmPFC

On the basis of previous behavioral and neuroimaging work, we expected that greater control would be required when predicting action based on a false versus true belief, or a desire to avoid versus approach an object. Behavioral data from the current study were consistent with these predictions. The neuroimaging results converge with the general executive literature in pinpointing dmPFC, comprising dorsal ACC and paracingulate gyrus, in supporting these more cognitively effortful scenarios (Botvinick et al., 1999, 2004; Bush et al., 2000). Factorial analysis (Table 1, Figure 3) showed that dmPFC was modulated by manipulating the content of specific ToM states. Investigation of the directional contrasts (Table 2) highlighted that these main effects were driven by those mental state concepts where the greatest need for control existed, such as false belief, avoidance, and unspecified desire. Notably, the novel, unspecified desire condition attracted the greatest increase in response latencies and made greater demands on dmPFC than both avoidance and approach desire reasoning. We suggest that this result is consistent with our suggestion that dmPFC serves conflict detection in support of control processes, because to predict the behavior of the agent with an unspecified desire, participants would have to withhold any response until they had determined what they thought the agent's preferences might be. Here, then, conflict exists not only between competing outcomes, such as the undesirable versus the desirable outcome, but also potentially between what the participant would do and what someone like the agent would do in that particular situation. Taken together, then, these data are further evidence that dmPFC serves a very general control function, with more specific functions—such as inhibition of self-perspective—supported by other neural regions.

ToM Reasoning and the rmPFC

Also of interest was the role of rmPFC in ToM. Existing literature, together with the task analysis presented here, suggests two possible roles for this region, and our task was designed to distinguish between them. First, the consistency with which rmPFC is recruited for ToM in previous research has led some authors to suggest that this region serves the function of representing mental states (Amodio & Frith, 2006; Frith & Frith, 2003, 2006). All conditions of our current paradigm required representation of the character's mental states, and so this interpretation of the role of rmPFC does not predict any variation in activation across conditions. Second, a growing literature indicates that thinking beyond the stimuli presented recruits rmPFC

Table 2. Directional Contrasts within the Factors of Belief and Desire

Region	*Hemi*	*Brodmann's area*	*MNI Coordinates*			*Z Value*
			x	*y*	*z*	
Belief						
B+ > B−						
Occipital pole	L	18	−20	−94	−10	5.52
Lateral occipital cortex, superior division	L	18	−24	−88	18	5.31
Lateral occipital cortex, inferior division	L	19	−36	−90	−12	4.02
B− > B+						
Temporoparietal junction	R	22	50	−50	26	7.18
Precuneus cortex	R	7	2	−66	48	5.87
Temporoparietal junction	L	40	−52	−52	30	5.76
Insular cortex	R	47	44	16	−6	5.63
Frontal orbital cortex	L	47	−32	26	−2	5.54
Supramarginal gyrus, posterior division	L	40	−44	−50	42	5.43
Precuneus cortex	L	7	−8	−64	50	5.38
Occipital pole	L/R	17	0	−92	−12	5.36
Middle frontal gyrus	R	44	50	20	38	5.31
Paracingulate gyrus	L	8	−4	20	48	5.21
Inferior frontal gyrus, pars opercularis	L	48	−54	16	0	5.20
Frontal pole	R	46	38	52	18	5.17
Middle frontal gyrus	L	9	−50	14	44	5.09
Paracingulate gyrus	R	32	2	42	28	5.00
Inferior frontal gyrus, pars triangularis	L	45	−60	22	8	4.96
Temporal pole	L	38	−52	16	−10	4.83
Frontal operculum cortex	L	47	−44	18	−4	4.71
Superior frontal gyrus	L/R	8	0	22	56	4.57
Superior frontal gyrus	R	9	2	40	42	4.53
Inferior frontal gyrus, pars opercularis	R	48	52	18	4	4.49
Lateral occipital cortex, superior division	R	39	44	−58	40	4.38
Supramarginal gyrus, anterior division	L	40	−54	−40	38	4.37
Lingual gyrus	R	18	4	−84	−16	3.97
Postcentral gyrus	L/R	5	0	−54	72	3.70
Desire						
D+ > D−						
Occipital pole	L	18	−22	−94	−8	8.00
Occipital pole	R	18	22	−96	−2	7.49
Lateral occipital cortex, inferior division	L	19	−36	−90	−12	5.60
Lateral occipital cortex, superior division	R	19	12	−86	44	5.24
Precentral gyrus	L	6	−54	−2	46	4.85
Precentral gyrus	R	6	50	−8	54	4.48

Table 2. (*continued*)

Region	*Hemi*	*Brodmann's area*	*MNI Coordinates*			*Z Value*
			x	*y*	*z*	
Postcentral gyrus	R	4	4	−36	56	4.26
Superior parietal cortex	R	5	20	−50	66	4.17
Postcentral gyrus	L	2	−28	−40	66	4.09
D+ > D±						
Occipital pole	L	18	−20	−96	−6	7.96
Occipital pole	R	18	22	−98	0	7.84
Occipital fusiform gyrus	R	18	20	−82	−10	6.38
Lingual gyrus	R	18	16	−88	−6	6.11
Occipital fusiform gyrus	L	18	−14	−84	−12	6.01
Intracalcarine cortex	R	17	14	−84	2	6.00
Precentral gyrus	L	6	−54	−2	46	5.27
Precentral gyrus	R	4	52	−4	38	4.61
Postcentral gyrus	R	3	24	−38	76	4.22
Postcentral gyrus	L	3	−24	−40	70	4.17
Superior parietal cortex	R	5	20	−50	68	4.09
Superior frontal gyrus	R	6	16	2	72	3.94
Superior temporal gyrus, anterior division	L	21	−56	2	−12	3.53
Planum polare	L	38	−58	2	−2	3.52
Central opercular cortex	L	48	−50	−2	8	3.25
D− > D+						
Paracingulate gyrus	R	8	2	24	48	8.06
Superior frontal gyrus	R	9	2	40	42	7.17
Superior frontal gyrus	L	8	−8	30	46	6.62
Frontal orbital cortex	R	47	34	22	−8	6.50
Middle frontal gyrus	R	45	52	28	24	6.48
Insular cortex	R	47	34	24	0	6.45
Temporoparietal junction	R	22	52	−56	26	6.39
Inferior frontal gyrus, pars opercularis	R	48	54	20	6	6.37
Frontal orbital cortex	L	47	−32	24	−8	6.08
Temporoparietal junction	L	40	−48	−52	42	5.61
Precuneus cortex	R	7	4	−68	42	5.35
Supramarginal gyrus, posterior division	R	40	52	−46	48	5.34
Supramarginal gyrus, posterior division	L	40	−46	−44	44	5.27
Precuneus cortex	L/R	7	0	−68	54	4.65
Lateral occipital cortex, superior division	R	39	44	−58	50	4.53
Precuneus cortex	L	7	−8	−64	50	4.15
Lateral occipital cortex, superior division	L	39	−50	−68	44	2.72
D− > D±						
Lateral occipital cortex, superior division	R	19	28	−80	24	4.43

Table 2. (*continued*)

Region	*Hemi*	*Brodmann's area*	*MNI Coordinates*			*Z Value*
			x	*y*	*z*	
Supramarginal gyrus, posterior division	L	22	−56	−46	10	4.42
Occipital pole	R	18	14	−88	22	4.39
Supramarginal gyrus, anterior division	L	40	−58	−38	32	4.36
Intracalcarine cortex	R	17	10	−70	14	4.29
Occipital pole	L	18	−18	−96	−2	4.14
Cuneal cortex	R	18	4	−84	32	4.11
Precuneus cortex	L	7	−4	−58	56	3.94
Superior temporal gyrus, posterior division	L	21	−54	−30	−2	3.89
Middle temporal gyrus, posterior division	L	21	−58	−24	−8	3.85
Lateral occipital cortex, inferior division	L	37	−46	−70	2	3.73
D± > D+						
Paracingulate gyrus	R	8	2	24	48	8.67
Paracingulate gyrus	L/R	8	0	28	42	8.48
Paracingulate gyrus	L	32	−4	34	36	8.31
Superior frontal gyrus	R	8	4	42	50	7.37
Superior frontal gyrus	L	8	−4	28	60	7.11
Insular cortex	R	47	36	22	−6	6.17
Lateral occipital cortex, superior division	L	39	−46	−60	36	5.63
Temporoparietal junction	L	39	−46	−58	42	5.29
Cingulate gyrus, posterior division	L	23	−2	−54	24	4.51
Lateral occipital cortex, superior division	R	22	60	−60	26	4.48
Temporoparietal junction	R	39	50	−58	28	4.37
Supramarginal gyrus, posterior division	L	40	−46	−50	52	3.83
Supramarginal gyrus, anterior division	R	2	54	−28	44	3.73
Supramarginal gyrus, anterior division	L	2	−46	−38	44	3.72
Precuneus cortex	R	7	2	−66	34	3.64
Supramarginal gyrus, posterior division	R	40	50	−44	54	3.52
Cingulate gyrus, posterior division	L/R	23	0	−16	28	3.48
Cingulate gyrus, anterior division	L/R	23	0	−10	28	3.21
Cingulate gyrus, posterior division	L	29	−4	−48	14	2.85
D± > D−						
Paracingulate gyrus	R	32	2	50	26	5.97
Paracingulate gyrus	L/R	32	0	38	34	5.97
Frontal pole	L/R	10	0	60	30	5.72
Frontal pole	L	9	−16	40	48	5.39
Superior frontal gyrus	L	8	−2	38	50	5.08

Table lists local maxima for cortical regions identified using a series of directional *t* contrasts, where $Z > 2.5$, $p_{corr} < .001$. All anatomically unique local maxima (with minimum peak separation of 5 mm) are listed. Brodmann's areas are approximate.

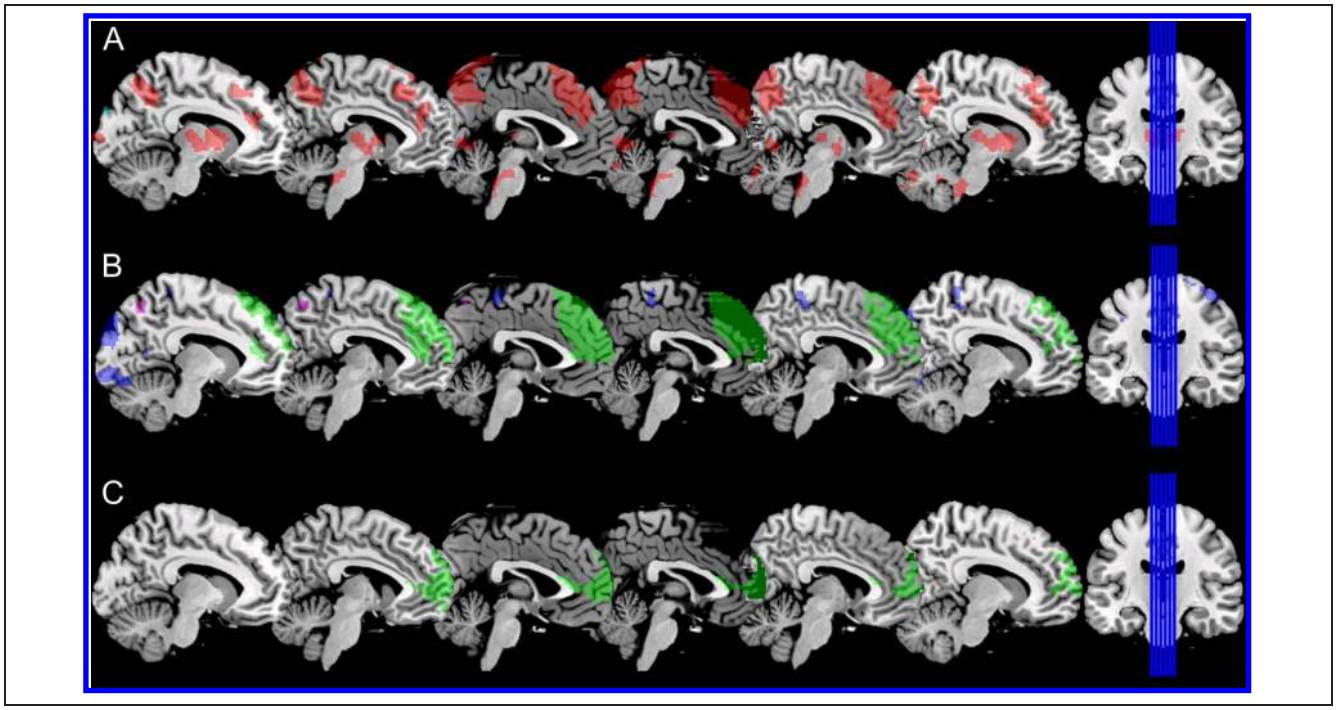

Figure 4. The group data are overlaid on the MNI brain template, showing significantly activated voxels where $Z > 2.5$, $p_{corr} < .001$. Slices from left to right, $x = -10, -6, -2, 2, 6, 10$, respectively. Maps reflect Z-corrected t-statistical images. (A) Voxels that are preferentially active during true versus false belief reasoning ($B+_{pref}$; cyan); false versus true belief reasoning ($B-_{pref}$; red). (B) Voxels that are preferentially active during approach versus unspecified AND avoidance desire ($D+_{pref}$; blue); avoidance versus approach AND unspecified desire ($D-_{pref}$; magenta); unspecified versus approach AND avoidance desire ($D\pm_{pref}$; green). (C) Voxels within the medial frontal cortex that are preferentially active for unspecified desire versus all other belief and desire conditions.

even in nonsocial contexts (Jenkins & Mitchell, 2009; Gilbert et al., 2007). Thinking beyond the stimuli and, in particular, so-called "abductive" inference to the best explanation is a frequent requirement of ToM, both in tasks and outside the laboratory. However, it is not a necessary feature, and it was not present in the belief factor of the current task, whereas the desire factor included one level that required abductive reasoning (D±) and two levels that only required deductive reasoning (D− and D+).

Consistent with Hartwright et al. (2012), factorial analysis identified that manipulating an agent's belief state did not modulate rmPFC (Table 1, Figure 3). Thus, there was no difference in how reasoning deductively about an agent with a true or false belief state was handled by this region. In contrast to this, manipulation of an agent's desire state was shown to modulate rmPFC. Note that this was not the case in our previous study, which did not require abductive reasoning in any condition. Directional and contrast masking analyses (Table 2, Figure 4) were used to clarify which of the variations in mentalizing was driving this effect. rmPFC was shown to respond preferentially when reasoning about an agent whose desire was unspecified (D±), over and above any of the other deductive belief and desire conditions (Figure 4C). Collectively, these data suggest that rmPFC is responsive to the requirement to reason abductively about mental states.

These findings converge with Jenkins and Mitchell (2009), who found that comprehension of a story whose causal structure was ambiguous or incomplete, rather than unambiguous and complete, preferentially recruited mPFC, including rmPFC. Such effects were found irrespective of whether the stories required inferences about a character's mental states, and indeed it is unclear in this study whether rmPFC was recruited for the ToM inferences themselves or just for general comprehension of an ambiguous context. The current study provides important clarity on this point by showing that rmPFC is indeed recruited for ToM inferences specifically in cases where abductive rather than deductive reasoning is required. In the broader social context, Van Overwalle (2009) notes that studies that invite richer inferences, such as trait ascription, recruit mPFC. Relatedly, Quadflieg et al. (2009) demonstrated that rmPFC is recruited when reasoning about the type of person (male/female/either) versus the type of place (indoors/outdoors/either) that is likely to be associated with an activity, such as mowing the lawn or watching talk shows. Thus, rmPFC was seen as an important neural substrate of the access and assignation of stereotype information. It is important to highlight, however, that this does not conflict with our assertion that rmPFC supports a general process that is engaged when reasoning abductively. A considerable literature demonstrates the automaticity of trait inferences and social

categorization (Greenwald & Banaji, 1995), for example, on the basis of an image of a face (Todorov, Said, Engell, & Oosterhof, 2008) or when primed subconsciously (Bargh, Chen, & Burrows, 1996). As such, all of our desire conditions featured the photographs of faces taken from a single database; our analyses would, therefore, subtract out those neural regions required for the attribution of stereotype schemas, as the potential for spontaneous trait ascription, including the automatic generation of stereotypes, is constant across all conditions. Our unspecified desire condition, on the other hand, is the only condition to require an abductive inference on the basis of such ascriptions. When considered alongside a literature that implicates rmPFC in autobiographical thinking, for example, in terms of imagining past or future events versus simply recalling such occurrences, prospection and the default mode network (see reviews by Schacter et al., 2012; Spreng et al., 2009), the commonality across these, and Jenkins and Mitchell (2009), is a shared process that reflects the assignation of information that is obtained through a rich, inferential process. The present paradigm varied the requirement for this process, by including a single, abductive reasoning condition alongside matched, a series of deductive reasoning conditions.

Cognitive versus Affective ToM

Qualitative reviews of the literature suggest a functional subdivision within mPFC, where a dorsal/rostral boundary may delineate cognitive versus affective ToM, respectively (Abu-Akel & Shamay-Tsoory, 2011; Carrington & Bailey, 2009; Lieberman, 2007; Amodio & Frith, 2006). Thus, belief reasoning would be expected to recruit more dorsal regions of mPFC, whereas desire reasoning would recruit rostral regions. While the current data might initially appear to favor this distinction, we suggest that a simple cognitive/affective division provides less explanatory power for our data than the task analysis proposed here.

First, this study suggests that it is likely to be the processing requirements within particular ToM concepts that modulate dmPFC (e.g., true versus false belief), rather than the cognitive or affective nature of the ToM concept. Our data identify that cognitively effortful situations involving false belief, avoidance, or unspecified desire reasoning make greater demands on dmPFC than less effortful ToM situations such as true belief or approach desire. We suggest that this effort, seen in increased response latencies and errors, is a reflection of increased conflict between alternative predictions for the agent. Thus, increased effort is associated with increased demand on dmPFC, regardless of the type of mental state being represented.

Second, while only our affective (desire) condition recruited rmPFC, this region was preferentially engaged as a function of the reasoning demands within this condition, rather than the mere requirement to infer desires. Specifically, a context that required abductive inference about desire was associated with increased demand on rmPFC, compared with conditions that only required deductive inferences about desire. Our data show that rmPFC is brought in to serve context-specific reasoning processes, such as when mentalizing beyond the information presented is required.

Representing Mental States

By definition, ToM requires people to hold in mind representations of mental states, and questions about this representational aspect of ToM have dominated thinking in the developmental, cognitive, comparative, and neuroscience literatures (Call & Tomasello, 2008; Saxe & Powell, 2006; Fodor, 1992; Leslie, 1987). However, identification of the neural basis of such representations has proved a surprisingly elusive target, with ongoing debates about the relative specificity of mPFC versus TPJ for such representations (Aichhorn et al., 2009; Scholz et al., 2009; Amodio & Frith, 2006; Saxe & Powell, 2006; Saxe & Wexler, 2005; Saxe & Kanwisher, 2003). This study was not designed as a strong test of the neural correlates of representing mental states, as we did not include conditions without mental states for comparison. However, the current findings do add to a growing body of evidence, suggesting that the mere representation of mental states is only part of the neurocognitive basis of ToM, in two important ways. First, other functional processes for cognitive control and reasoning are integral to ToM and recruit neural regions supporting these processes in ways that can be predicted from functional analysis of ToM tasks. Second, even if a consensus does emerge on neural regions that are involved in representing mental states, it seems unlikely that this function will be sufficient to explain patterns of activity in those neural regions during ToM tasks. In this study, mental states needed to be represented in all conditions, and yet we observed condition-wise variation in activity in the neural regions most often suggested to be the neural basis of representing mental states (rmPFC and bilateral TPJ). Such variation can be understood by appeal to other functional aspects of ToM, such as the need for cognitive control, and the need for different kinds of reasoning. We suggest that this makes vivid the suggestion that ToM is subserved by a network, which may be comprised of distinct functional and anatomical components, but whose activity can only be understood by considering the network as a whole and the tasks in which it is engaged.

Acknowledgments

This research was supported by the Economic and Social Research Council, award number ES/G01258X/1.

Reprint requests should be sent to Charlotte E. Hartwright, School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, or via e-mail: cee849@bham.ac.uk.

REFERENCES

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia, 49,* 2971–2984.

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience, 21,* 1179–1192.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277.

Apperly, I. (2010). *Mindreaders: The cognitive basis of theory of mind*. Hove: Psychology Press.

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental continuity in theory of mind: Speed and accuracy of belief-desire reasoning in children and adults. *Child Development, 82,* 1691–1703.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230–244.

Botvinick, M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8,* 539–546.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402,* 179–181.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4,* 215–222.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12,* 187–192.

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping, 30,* 2313–2335.

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *Neuroimage, 60,* 1652–1661.

Fodor, J. A. (1992). A theory of the childs theory of mind. *Cognition, 44,* 283–296.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50,* 531–534.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society, Series B, Biological Sciences, 358,* 459–473.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind". *Trends in Cognitive Sciences, 7,* 77–83.

German, T., & Hehman, J. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition, 101,* 129–152.

Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D., et al. (2006). Functional specialization within rostral prefrontal cortex (area 10): A meta-analysis. *Journal of Cognitive Neuroscience, 18,* 932–948.

Gilbert, S. J., Williamson, I. D. M., Dumontheil, I., Simons, J. S., Frith, C. D., & Burgess, P. W. (2007). Distinct regions of medial rostral prefrontal cortex supporting social and nonsocial functions. *Social Cognitive and Affective Neuroscience, 2,* 217–226.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition—Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *Neuroimage, 61,* 921–930.

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20,* 404–410.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24,* 1377–1388.

Leslie, A. M. (1987). Pretense and representation—The origins of theory of mind. *Psychological Review, 94,* 412–426.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58,* 259–289.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62,* 103–134.

Menzies, T. (1996). Applications of abduction: Knowledge-level modelling. *International Journal of Human-Computer Studies, 45,* 305–335.

Morris, H. C. (1992). Logical creativity. *Theory & Psychology, 2,* 89–107.

Pagnucco, M. (1996). *The role of abductive reasoning within the process of belief revision*. Unpublished doctoral dissertation, University of Sydney, Australia.

Prado, J., Chadha, A., & Booth, J. R. (2011). The brain network for deductive reasoning: A quantitative meta-analysis of 28 neuroimaging studies. *Journal of Cognitive Neuroscience, 23,* 3483–3497.

Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience, 21,* 1560–1570.

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience, 5,* 184–194.

Rothmayr, C., Sodian, B., Hajak, G., Döhnel, K., Meinhardt, J., & Sommer, M. (2011). Common and distinct neural networks for false belief reasoning and inhibitory control. *Neuroimage, 56,* 1705–1713.

Rushworth, M. F. S., Buckley, M. J., Behrens, T. E. J., Walton, M. E., & Bannerman, D. M. (2007). Functional organization of the medial frontal cortex. *Current Opinion in Neurobiology, 17,* 220–227.

Saxe, R., & Andrews-Hanna, J. R. (n.d.). Retrieved October 7, 2012, from saxelab.mit.edu/stimuli.php.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *Neuroimage, 19,* 1835–1842.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17,* 692–699.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43,* 1391–1399.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron, 76,* 677–694.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal

junction are selective for theory of mind and exogenous attention. *PLoS One, 4,* e4869.
Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *Neuroimage, 35,* 1378–1384.
Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21,* 489–510.
Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12,* 455–460.
Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30,* 829–858.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103–128.

This article has been cited by:

1. John Haracz. 2019. Neuroeconomics of Asset-Price Bubbles: Neuroimaging and Digital Technology for the Prediction and Prevention of Major Bubbles. *SSRN Electronic Journal* . [Crossref]
2. . The Social Brain in Adolescence and Adulthood: Lessons in Mindreading 114-146. [Crossref]
3. Serge A. Mitelman, Marie-Cecile Bralet, M. Mehmet Haznedar, Eric Hollander, Lina Shihabuddin, Erin A. Hazlett, Monte S. Buchsbaum. 2018. Positron emission tomography assessment of cerebral glucose metabolic rates in autism spectrum disorder and schizophrenia. *Brain Imaging and Behavior* **12**:2, 532-546. [Crossref]
4. Yong-Ku Kim, Ho-Kyoung Yoon. 2018. Common and distinct brain networks underlying panic and social anxiety disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **80**, 115-122. [Crossref]
5. Paul F. Hill, Richard Yi, R. Nathan Spreng, Rachel A. Diana. 2017. Neural congruence between intertemporal and interpersonal self-control: Evidence from delay and social discounting. *NeuroImage* **162**, 186-198. [Crossref]
6. Bradley D. Mattan, Kimberly A. Quinn, Stephanie L. Acaster, Rebecca M. Jennings, Pia Rotshtein. 2017. Prioritization of Self-Relevant Perspectives in Ageing. *Quarterly Journal of Experimental Psychology* **70**:6, 1033-1052. [Crossref]
7. Bradley D. Mattan, Pia Rotshtein, Kimberly A. Quinn. 2016. Empathy and visual perspective-taking performance. *Cognitive Neuroscience* **7**:1-4, 170-181. [Crossref]
8. Charlotte E. Hartwright, Robert M. Hardwick, Ian A. Apperly, Peter C. Hansen. 2016. Resting state morphology predicts the effect of theta burst stimulation in false belief reasoning. *Human Brain Mapping* **37**:10, 3502-3514. [Crossref]
9. Benjamin O. Turner, Nicole Marinsek, Emily Ryhal, Michael B. Miller. 2015. Hemispheric lateralization in reasoning. *Annals of the New York Academy of Sciences* **1359**:1, 47-64. [Crossref]
10. Yin Wang, Susanne Quadflieg. 2015. In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience* **10**:11, 1515-1524. [Crossref]
11. Dana Samson, Sarah Houthuys, Glyn W. Humphreys. 2015. Self-perspective inhibition deficits cannot be explained by general executive control difficulties. *Cortex* **70**, 189-201. [Crossref]
12. Alexander Otti, Afra M. Wohlschlaeger, Michael Noll-Hussong. 2015. Is the Medial Prefrontal Cortex Necessary for Theory of Mind?. *PLOS ONE* **10**:8, e0135912. [Crossref]
13. Arjen Stolk, Daniela D'Imperio, Giuseppe di Pellegrino, Ivan Toni. 2015. Altered Communicative Decisions following Ventromedial Prefrontal Lesions. *Current Biology* **25**:11, 1469-1474. [Crossref]
14. Sara M. Schaafsma, Donald W. Pfaff, Robert P. Spunt, Ralph Adolphs. 2015. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences* **19**:2, 65-72. [Crossref]
15. Tobias Schuwerk, Martin Schecklmann, Berthold Langguth, Katrin Döhnel, Beate Sodian, Monika Sommer. 2014. Inhibiting the posterior medial prefrontal cortex by rTMS decreases the discrepancy between self and other in Theory of Mind reasoning. *Behavioural Brain Research* **274**, 312-318. [Crossref]